Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle *Since 1965*

Deep Learning on Multilingual Social Media

CHALLENGE

Real-time situational awareness, anomaly detection, and event summarization in social media require understanding communication across languages, media, and platforms. Predictive analytics around human communication requires that real-time data streams can be reasoned against in a scalable fashion. Communication on social media involves evolving vocabularies, user relationships, and topics. The techniques used for such analyses have to be continually updated, handle the streaming nature of data, work at web scale across many languages and geolocations, and account for evolving social network structure.

The objective of this project is to build neural network models (Figure 1) for understanding human communication in social media. These models will learn high-quality multilingual, network structure-aware and time-aware distributed (e.g., across geo-locations) representations from social media data on an unprecedented scale, including words, concepts, hashtags, mentions, entities, and events (attack, protests, natural disasters, etc.).

CURRENT PRACTICE

The existing approaches to large-scale real-time text analysis are keyword-based and rely on topic modeling or trend detection that lacks deep semantic understanding of word dynamics in context. Recent advances in distributional semantics and neural networks allow projecting language into low-dimensional space to find

Enabling a language-independent view of global events and trends in real time.

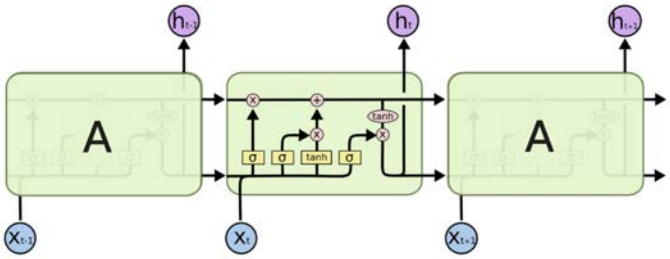

Figure 1: Recurrent neural networks allow learning over streaming data.

similarities among semantically similar words. Learning embeddings from streaming social network data across many languages and geolocations is challenging because of noise, sparsity, data drift, dynamic nature, and other issues related to social media.

TECHNICAL APPROACH

Compositional semantics states that content in similar context has similar meaning. Context might change with time and in space and may depend on broader contexts, such as social network structure. We propose to consider a broader definition of context by relying on a social network structure and adding temporal dependency while

learning distributed word representations from streaming social media data across multiple languages jointly.

For example, this approach will allow us to find semantic relatedness among evolving concepts in social media: Belgium and bombing (Figure 2) and terrorism and Islamic (Figure 3) for the time period between March 15 and 29, 2016. To learn distributed representations, we will rely on deep neural network architectures. However, unlike other approaches, this method will learn latent embeddings in the following unique ways:

- Jointly across many languages
- On a large amount of social media data
- In a streaming fashion
- On a user level rather than a tweet or document level.

IMPACT

By incorporating language, time, space, user information, and a social network evolution history, our work will enable downstream forecasting and prediction—question answering, paraphrase detection, event recognition, and other predictive analytics—impossible or impractical with previous text analysis strategies. Moreover, our models will enable a language-independent view of global events and trends in real time.

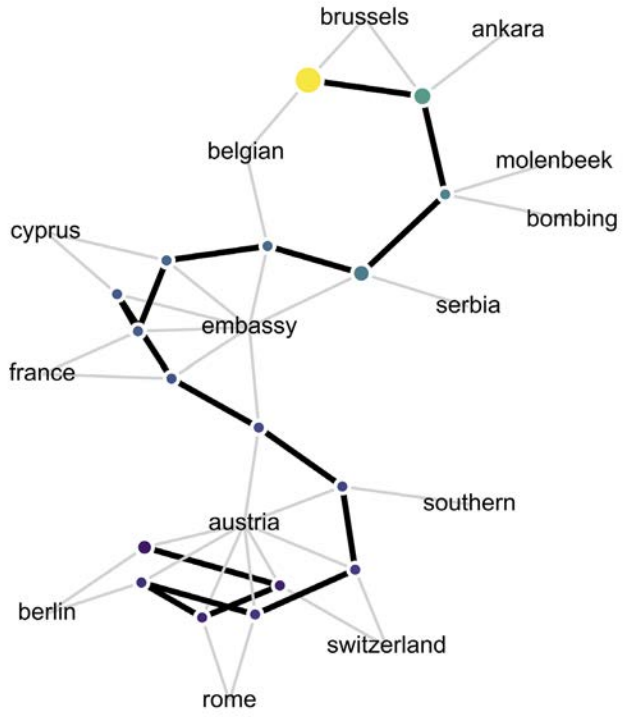

Figure 2: Concepts related to "Belgium" on Twitter.

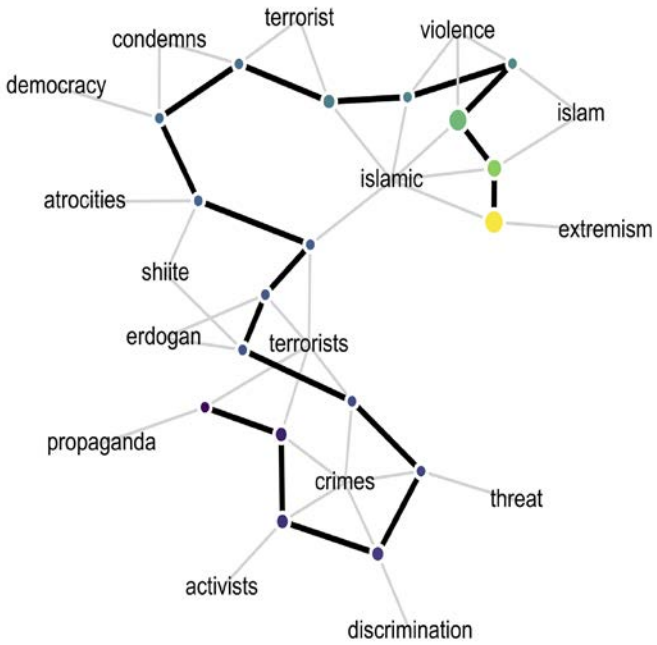

Figure 3: Concepts related to "terrorism" on Twitter.

Contacts

Svitlana Volkova
Principal Investigator
(509) 372-6585
svitlana.volkova@pnnl.gov

John R. Johnson
Program Director
(509) 375-2651
John.Johnson@pnnl.gov

Nathan Hodas
Principal Investigator
(509) 375-2862
nathan.hodas@pnnl.gov

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965